ResearchOnline@JCU

Steinig, Eike J., Neuditschko, Markus, Khatkar, Mehar S., Raadsma, Herman W., and Zenger, Kyall R. (2016) *NETVIEW P: a network visualization tool to unravel complex population structure using genome-wide SNPs.* Molecular Ecology Resources, 16 (1). pp. 216-227.

http://dx.doi.org/ 10.1111/1755-0998.12442

NetView P: A network visualization tool to unravel complex population structure using genome-wide SNPs

Eike J. Steinig[1], Markus Neuditschko[2], Mehar S. Khatkar[2,3], Herman W. Raadsma[1,2,3] and Kyall R. Zenger[1,3]

1. College of Marine and Environmental Sciences, James Cook University, Townsville, Queensland, Australia
2. Reprogen – Animal Bioscience, Faculty of Veterinary Science, University of Sydney, Camden, New South Wales, Australia
3. Centre for Sustainable Tropical Fisheries and Aquaculture, Townsville, Queensland, Australia

Keywords: Netview, population genetics, network analysis, graph theory, wild and captive populations, SNP

Corresponding author: Eike J. Steinig, CMES James Cook University, Townsville, Australia

eikejoachim.steinig@my.jcu.edu.au

Running title: NetView P: Network Visualization for Python

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12442

Abstract

Network-based approaches are emerging as valuable tools for the analysis of complex genetic structure in both wild and captive populations. *NetView P* combines data quality control with the construction of population networks based on mutual *k*-nearest-neighbours thresholds applied to genome-wide SNPs. The program is cross-platform compatible, open-source and efficiently operates on data ranging from hundreds to hundreds of thousands of SNPs through multiprocessing in Python. We used the pipeline for the analysis of pedigree data from simulated (n = 750, SNPs = 1279) and captive Silver-lipped Pearl Oysters (n = 415, SNPs = 1107), wild populations of the European Hake from the Atlantic and Mediterranean (n = 834, SNPs = 380) and Gray Wolves from North America (n = 239, SNPs = 86,103). The population networks effectively visualize large- and fine-scale genetic structure within and between populations, including family-level structure and relationships. *NetView P* comprises a network-based addition to other population analysis tools and provides user-friendly access to a complex network analysis pipeline through implementation in Python.

Introduction

The interaction of evolutionary forces such as genetic drift, gene flow, natural and artificial selection gives rise to genetic structure within and between populations. Revealing the extent of such structure can provide valuable insights into the history and evolutionary trajectory of natural populations (Holsinger & Weir 2009; Frean *et al.* 2013) and is frequently used to correct population stratification in genome-wide association studies (Price *et al.* 2006, 2010). An understanding of population structure is also crucial for the management of genetic resources in conservation and breeding programs. For instance, in natural populations neutral and adaptive differentiation of populations can be used for the evaluation and delineation of conservation management units (Crandall *et al.* 2000; Palsbøll *et al.* 2007; Funk *et al.* 2012). In captive populations, family assignments and relationships are highly relevant to the avoidance of inbreeding and choosing individuals for selective breeding programs (Hayes *et al.* 2006; Lind *et al.* 2009, 2012).

In the past decade, innovations in genome-wide sequencing methods and analytical tools have allowed for the recovery of large-scale genomic information, particularly with respect to non-model species (Miller *et al.* 2007; Baird *et al.* 2008; Peterson *et al.* 2012; Catchen *et al.* 2013; Ellegren 2014). However, in both natural and captive populations, the final analysis of such data critically depends on methods that can effectively reveal and visualize genetic structure within and between populations. A variety of approaches are commonly employed for this purpose, such as STRUCTURE (Pritchard *et al.* 2000) or Discriminant Analysis of Principal Components (DAPC) (Jombart *et al.* 2010). Model-based analyses are rooted in population genetic theory, but can depend on complex statistical approaches and make stringent assumptions on the data. Non-parametric approaches usually make fewer assumptions, but it can be difficult to visualize populations in a reduced dimensional space, especially in data sets that consist of closely related sub-populations (Neuditschko *et al.* 2012).

Network theory is emerging as a promising alternative to traditional methods of population analysis (Dyer & Nason 2004; Rozenfeld *et al.* 2008; Moalic *et al.* 2011; Neuditschko *et al.* 2012; Kivelä *et al.* 2014). In the graph-theoretical approach, the data is depicted by the topology of a network. In a population genetics context, multi-locus genotype data can be used to construct networks of individuals (Moalic *et al.* 2011; Neuditschko *et al.* 2012) or pre-determined populations (Dyer & Nason 2004; Rozenfeld *et al.* 2008; Noutsos *et al.* 2014). Network analyses lack many of the assumptions of model-based approaches, such as underlying population models or prior geographical clustering (Kivelä *et al.* 2014) and can depict connectivity and information flow within and between populations (Rozenfeld *et al.* 2008; Neuditschko *et al.* 2012; Kivelä *et al.* 2014). Their use for population analysis has been exemplified in the desert cactus *Lophocereus schottii* (Dyer & Nason 2004), the metapopulation system of the seagrass *Posidonia oceania* (Rozenfeld *et al.* 2008), a hybridization study of two micro-algae of the genus *Fucus* (Moalic *et al.* 2011) and diverging species of the plant *Aquilegia* (Noutsos *et al.* 2014).

In contrast to these studies, the network analysis and visualisation pipeline NetView (Neuditschko *et al.* 2012) was developed for genome-wide data and was successfully applied to bovine and human populations including hundreds of individuals and tens of thousands to millions of SNPs (Neuditschko

et al. 2012). NetView is currently implemented as a manually operated pipeline, which is largely inefficient for the analysis of multiple data sets or parameter values and does not facilitate access for the wider community. Furthermore, its initial application was based on relatively divergent populations (Neuditschko *et al.* 2012). However, wild populations display complex genetic structure (Lind *et al.* 2007; Arnaud-Haond *et al.* 2008; Milano *et al.* 2014; Pujolar *et al.* 2014; Cronin *et al.* 2015) and captive breeding programs may depend on the identification of closely-related individuals and families in successive generations (Lind *et al.* 2012).

In this study, we developed a comprehensive implementation of NetView in Python. *NetView P* is cross-platform compatible, open-source and supports multiprocessing for efficient data analysis over a wide range of parameters. We demonstrate the application of the program on simulated and empirical pedigrees of the Silver-lipped Pearl Oyster (*Pinctada maxima*) from Indonesia (Jones *et al.* 2013b; a), wild populations of the European Hake from the Atlantic and Mediterranean (*Merluccius merluccius*, Milano *et al.* 2014) and divergent populations of Gray Wolves from North America (*Canis lupus*, Cronin *et al.* 2015). We show that the networks can effectively visualize population- or family-level assemblages and relationships. Sample sizes ranged from below a hundred to several hundred individuals, genotyped at hundreds to tens of thousands of SNPs.

Materials and Methods

NetView P

NetView P connects the general components of network-based analysis pipelines described in NetView: (i) quality control of raw data, (ii) computation of a genetic distance matrix, (iii) construction of a network, (iv) detection of community structure (optional) and (v) visualisation of the final network topology (Neuditschko *et al.* 2012). As in the original implementation of NetView, the initial quality control and calculation of a shared-allele distance matrix (1-IBS) are carried out in PLINK v1.07(Purcell *et al.* 2007; Neuditschko *et al.* 2012). However, network construction and clustering were originally implemented manually through the super-paramagnetic clustering algorithm SPC (Blatt *et al.* 1996; Barad 2003) in the software Sorting Points Into Neighbourhoods (SPIN)

(Tsafrir *et al.* 2005). In order to develop an open-source, cross-platform version of NetView in Python, the SPC and SPIN were omitted from *NetView P*. Instead, the initial network construction using mutual *k*-nearest-neighbour (mk-NN) thresholds was adopted from Barad (2003). The resulting components of the graph can optionally be connected through a minimum spanning tree (MST) determined by Prim's algorithm (Prim 1957). More precisely, given a symmetrical pairwise distance matrix X and the number of nearest neighbours k, we first construct $G_{mut}\ (n, k)$ where individuals (nodes) X_i and X_j are connected by an undirected edge E_{ij} if $X_i \in k\text{-}NN(X_j)$ and $X_j \in k\text{-}NN(X_i)$ (Maier *et al.* 2007). In order to recover a connected network, the undirected edges of the minimum spanning tree associated with X are added to $Gmut$ if $Eij \notin Gmut$. The weight of each edge is assigned the genetic distance between Xi and Xj.

A connected graph is sometimes required for downstream detection of community structure, such as with SPC or Infomap (Rosvall & Bergstrom 2008). However, edges derived from the MST can affect the positioning of nodes that link cluster previously not connected through mutual k-NN. We therefore included additional edge colouration in the final network files in order to gauge their effect on the placement of connecting nodes. It should be noted that the construction of a connected network is optional and can be switched off, providing the user with a clear representation of individual, communities of samples at a particular value of *k*. The open-source, information-theoretic community-detection algorithm Infomap (Rosvall & Bergstrom 2008) was implemented as a replacement for SPC. However, like other network-based methods for population analysis we anchor our interpretations in the network topologies, rather than community structure (Dyer & Nason 2004; Rozenfeld *et al.* 2008; Moalic *et al.* 2011; Kivelä *et al.* 2014). Finally, it should be noted that the network construction is independent of prior information and based on the genetic distances between individuals, unlike methods that require pre-determined populations such as F_{ST} or Discriminant Analysis of Principal Components (DAPC) (Jombart *et al.* 2010).

The network topologies are dependent on a single user-defined threshold parameter, the number of mutual nearest-neighbours (*k*). There is currently no appropriate optimisation for *k* (Neuditschko *et al.* 2012), but the effect of the parameter on the connectivity of the networks offers an intriguing

possibility to investigate population structure at different levels of genetic similarity, alternatively focusing on fine-scale structure (connecting fewer, more closely related samples at small *k*) or large-scale patterns of admixture (connecting more distantly related samples at large *k*) (Neuditschko *et al.* 2012). As suggested by Neuditschko *et al.* (2012), an appropriate, empirical value is $k = 10$. Nevertheless, the networks should generally be explored within a reasonable range of the parameter (e.g. $k = 5 - 40$) and a stepwise reduction in *k* is recommended for small sample collections (Neuditschko *et al.* 2012). It should be noted that the application of the mutual nearest-neighbour threshold is not based on established population models, but rather uses a simple machine learning algorithm on a given similarity matrix. However, we will show with simulated and empirical data that it can effectively recover and visualize population structures, including family-level assemblages and relationships.

The general workflow of the pipeline is outlined in Figure 1. The computational implementation is open-source and cross-platform compatible through Python. Parameters and options can be specified through a command line version or a simple, user-friendly GUI. Input formats are the PED/MAP for PLINK (directly compatible with STACKS, Catchen *et al.* 2013) or a simple matrix of SNPs. Alternatively, a pre-computed symmetrical distance matrix can be specified, which allows the user to implement their preferred quality control parameters or distance measures. This makes the pipeline applicable to any data from which such a distance matrix can be calculated, e.g. for the study of biogeographical provinces (Moalic *et al.* 2012). Connected networks can be constructed by including edges from the MST. Node colours and shapes are generated according to an additional attribute file (e.g. specifying colours for sampling site, population, sex, phenotype or pedigree) or automatically derived from the clustering results of Infomap. The final network files are formatted as edge lists and can be loaded into compatible visualisation platform such as Pajek (Batagelj & Mrvar 1998), iGraph (Csardi & Nepusz 2006), Gephi (Bastian *et al.* 2009) or Cytoscape (Smoot *et al.* 2011).

Simulated Data

In order to evaluate the capacity of the pipeline to detect family-level assemblages and relationships through successive generations, we simulated a data set of SNPs based on population parameters of *P. maxima* using QMSim (Sargolzaei & Schenkel 2009). The initial founder generation was the last of 1000 historic simulations containing 430 individuals each, equal to the effective population size of wild *P. maxima* (Lind *et al.* 2007). From this founder population, 20 males and 20 females were used for breeding, each mating producing 50 offspring. The genome map is comparable to the oyster linkage map from (Jones *et al.* 2013b), comprising 14 chromosomes with 4200 SNPs placed proportional to the chromosomes lengths. The simulation was run for 10 discrete generations with random selection of parents and genotypes provided for the last three generations (F8, F9, F10). For demonstration, we reduced the final dataset (n = 3000, SNPs = 4200) to three randomly selected families of F10 and included their parental families from the previous generations F8 and F9.

Empirical Data

We assembled three empirical data sets of wild and captive populations, comprising variable numbers of samples and SNPs. The first data set is comparable to the simulated dataset for *P. maxima* and contained the geographical origin or ancestry of samples, pedigree records and 1,147 EST-derived single nucleotide polymorphism (SNP) markers previously developed by(Jones *et al.* 2013b; a). SNPs were derived from farmed oysters over two consecutive generations (F_1 and F_2), initially founded from three natural populations (F_0, Aru Islands, Bali, West-Papua) and reared at two commercial sites in Indonesia. Pedigree records identified a total of twenty-nine putative, heterogeneously sized families, twelve in F_1 and seventeen in F_2. These were labelled with their respective generation, ancestry line of dame and sire, and a family number, if multiple ancestry combinations were present (A = Aru, B = Bali, W = West-Papua, U = Unknown; e.g. F1-UW-1 or F2-AW-2). For visual simplicity and complete validation of the recovered pedigree structure, we only included families for which both parents were available and had been genotyped. This yielded a total of fourteen families: three in in F_0, ten in F_1 and three in F_2.

The second data set contained was derived from a study on the European Hake (*M. merluccius*) (Milano *et al.* 2014) and included wild fish from nineteen geographically distinct sites within the Mediterranean and Atlantic, genotyped at 380 EST-derived SNPs (pairwise F_{ST} = 0.004 - 0.028). The third data set comprised eight variably differentiated, wild populations of the Gray Wolf (*Canis lupus*) from North America (mean F_{ST} = 0.0342 - 0.3448), genotyped at 123,801 SNPs (Illumina 170K CanineBeadChip) including three game management units (GMUs) (Cronin *et al.* 2015). Additional network files for the wild founder populations and the full pedigree of simulated and captive *P. maxima* can be found in the Data Availability section and Supporting Materials (S1). In general, we expected the networks to recover population structures as determined for neutral markers by Milano *et al.* (2014) and recover the diverse relationships between wolf populations by Cronin *et al.* (2015). We also expected the networks to accurately depict the relationships (parents, half-siblings) between families and generations, as determined by the simulated and external pedigree records for *P. maxima*.

Network Construction

In the computational implementation, all data sets were first subjected to quality control using PLINK v1.07 (Purcell *et al.* 2007). Samples and SNPs were excluded based on frequency of missing data (> 0.1), minor allele frequency (< 0.01) and significant derivation from Hardy-Weinberg equilibrium (P < 0.001), as recommended by Neuditschko *et al.* (2012). Quality control produced the final data sets for the simulation (n = 750, SNPs = 1279), captive *P. maxima* (n = 415, SNPs = 1107), *M. merluccius* (n = 834, SNPs = 380) and *C. lupus* (n = 239, SNPs = 86103). A shared-allele distance matrix (1-IBS) was then calculated for each data set in PLINK v1.07. Connected networks were constructed at k = 10 (with MST), which captured both fine- and large-scale genetic structure as suggested in the original implementation by Neuditschko *et al.* (2012). In addition, the simulated data was constructed without edges from the MST. The final network visualisations were based on the organic and circular layouts from Cytoscape (Neuditschko *et al.* 2012). The NetworkAnalyzer (Doncheva *et al.* 2012) plugin was used for *M. merluccius* and *C. lupus* to generate the degree centrality for each node (the number of direct connections to other nodes, here proportional to node size) which has been used to distinguish 'unrelated' individuals in population networks (Neuditschko *et al.* 2012). A brief discussion on the

effect of *k* on the network topologies, as well as the application of the Infomap clustering and a comparison of results from PCA/DAPC for *P. maxima* can be found in the Supporting Material (S2-S4).

Results

European Hake and Gray Wolves

The network of the European Hake (Figure 1) agreed with structure detected in putatively neutral markers by Milano *et al.* (2014). The weak, but statistically significant break between populations from the Atlantic and the Mediterranean was clearly visualized in the network topology and the separation in the network corresponded to the genetic discontinuity observed in the Eastern Atlantic and the Alboran Sea (Milano *et al.* 2014). The networks showed relatedness of some samples from the southern Atlantic with samples from the Mediterranean, particularly from Algeria. These results also support observations of a higher rate of genetic contribution from individuals in the Atlantic to individuals from Algeria (see orange nodes in Figure 1) (Milano *et al.* 2014).

The networks of the Gray Wolves showed a division into several distinct populations (Figure 3). The Great Lakes population from Montana and the population from New Mexico appeared the most divergent, indicated by their isolated location and single edges (MST) between individuals linked to their respectively proximate populations. The US Northern Rocky Mountains population (Idaho, Minnesota, and Wyoming) formed an admixed cluster and included some individuals from British Colombia. Wolves from British Colombia linked closely with one part of the population in mainland Southeast Alaska (GMU1C), which in turn linked with wolves in GMU1A and GMU1B. The latter two units also showed a close relationship with the island-based GMU2 and GMU3, which were otherwise clearly differentiated in the networks. Further sub-structure within both Interior and Southeast Alaska (GMU2) was evident. The interior population of Alaska appeared more divergent from Southeast Alaska, but demonstrated linkage to some individuals from the Rocky Mountains and British Colombia. A single individual representing GMU1D was located near Interior Alaska. Lastly,

several distinct outliers could be detected by their low degree centralities and single edges removing them from the main communities, including one anomalous individual from Idaho linked with New Mexico.

Silver-lipped Pearl Oyster

Simulated

We first constructed a visualization of the simulated dataset without edges derived from the MST (Figure 4A). In this construction, assignment of individuals to their respective families is nearly complete, with each generation ordered manually and highlighted with different node colours (F8: purple, F9: green, F10: orange). The exceptions were two genetically distinct individuals in F10-FAM44, visibly separated from their native family. Although this network accurately determines the family clusters and assignments as simulated in the pedigree (Figure 4B), it does not depict information on the general relationships between the families. If we include the edges of the MST (red edges, Figure 5) the network visualization starts to reflect the relationship between the families and generations. For instance, on a fine-scale, parents (dark red nodes) of F9 and in some cases of F10 (F10-FAM43, -FAM48) are connected to their progeny family by the edges of the MST. Sometimes parents are drawn away from their families (e.g. F8-FAM1-7550 or F9-FAM30-8504) and are clearly visible in connecting successive generations. This parentage assignment was limited in F10, as for instance demonstrated by two parents in F9-FAM39, which did not connect to their respective offspring families F10-FAM44 and -FAM48. However, the general pedigree structures can be discerned (cf. Figure 4B), with the lower part of the graph from F8-FAM8 representing the lineage L2 with two F10 families F10-FAM43 and -FAM48 nested within their progenitor families in F8 and F9. This was also the case for the families from L1, which were connected by MST edges according to the simulated pedigree (cf. Figure 3B). Although the designation into generations and parentage assignment may be more complex in the absence of prior information, the networks accurately assign individuals to their respective families and with additional information such as known pedigree and MST edges, accurately reflect inter-generational and parentage relationships.

Empirical

The network of three generations of captive oysters recovered the family-level population structure mostly as expected from the external pedigree records (Figure 6). All families in F_1 and F_2 could be accurately distinguished, including half-sibling relationships between F1-BB-1 and F1-BB2, as well as F1-BB-3 and F1-BB-4 (asterisk, Figure 6). Parents from the three founder populations (triangular nodes) were located within or in the immediate vicinity of their offspring in F_1. However, for parents in the second generation (rectangular nodes) parental assignment was limited to F2-WW-1, with one parent of F2-BW-1 and F2-BW3 retained in F1-WW-1. Genetically distinct individuals belonging to the families were also recognisable, as seen in two samples of F0-ARU near F0-BAL and F0-ARU, and two samples in F1-WW-1. In the founder generation, the two more closely related populations from BAL and WPA (Lind *et al.* 2007) clustered together, which was likely facilitated by a relatively small number of samples from WPA. Admixture between ARU and BAL was also evident, largely corresponding to the wild population structure of *P. maxima* from Indonesia (see Supplementary Material S1). All in all, the genetic structure determined from previously assembled pedigrees was accurately visualized in the networks, representing both family-level relationships and parental assignments in F_1 and (partially) in F_2.

Discussion

NetView P is a comprehensive, cross-platform and open-source computational implementation of the original NetView (Neuditschko *et al.* 2012). The pipeline is based on the construction of a population network using mutual *k*-nearest-neighbours thresholds on a genetic distance matrix calculated from genome-wide SNPs. *NetView P* is applicable to commonly encountered data from both natural and captive populations in diverse population settings. The networks constructed in this study revealed large- and fine-scale patterns of population structure within and between closely related families over three consecutive generations in the Silver-lipped Pearl Oyster, the oceanic divide of the European Hake in the Atlantic and Mediterranean and patterns of differentiation and admixture in wild populations of the Gray Wolf.

The structure of the wild populations corresponded to the results expected from the studies by Milano *et al.* (2014) and Cronin *et al.* (2015). For instance, the transplantation of several wolves from British Colombia into the Rocky Mountains populations from Idaho, Minnesota and Wyoming (Cronin *et al.* 2015) was evident in the networks, showing some wolves from British Colombia clustering distinctly within the admixed population from the Rocky Mountains. Furthermore, the discrete clustering of populations from New Mexico and Minnesota is consistent with previous data from genome-wide SNPs (vonHoldt *et al.* 2011; Cronin *et al.* 2015). The large-scale divide between the Atlantic and Mediterranean populations of the European Hake observed in the networks also corresponded to previous results (Milano *et al.* 2014), particularly considering a relatively low resolution from 380 predominantly neutral SNPs. Milano *et al.* (2014) discovered fine-scale structure in a small number of putatively selected SNPs, corresponding to regional divisions within the Atlantic and Mediterranean. However, the application of the pipeline to a very small number of markers, as is often the case for markers that are under possible selective pressures, remains to be investigated.

The networks of pearl oysters recovered the simulated and real-world families and family-level relationships over three consecutive generations, including founder parentage and half-siblings in the first generation. However, recognition of parentage was limited in the second generation, although parents were still drawn out from their native family clusters in the networks. Considering that routine hatchery practices often involve mass spawning and communal rearing of families, coupled with high fecundities and the expense of keeping detailed pedigree records (Lind *et al.* 2009, 2012), our study shows that retrospective genotyping of adult oysters could be used to determine the genetic structure of families in captive populations of *P. maxima*. Breeding strategies based on kinship information of individuals critically depend on the identification and accurate assignment of individuals to family groups (Russello & Amato 2004; Ivy & Lacy 2012; Lind *et al.* 2012). The ability to reveal fine-scale genetic relationships between individuals and families could be applicable to commercial or conservation breeding programs, particularly where no prior pedigree information is available. In addition to the validation of the networks by previously collected pedigree records, our baseline

simulation further supports the capacity to recover family-level structure and relationships with *NetView P*.

Overall, the pipeline was effective for detecting both large- and fine-scale genetic structure at $k = 10$. The performance of the method has also been shown to be adequate when employing a small number of samples per population ($n < 20$) combined with a step-wise reduction of k (Neuditschko *et al.* 2012). Nevertheless, the networks should be investigated within a reasonable range of the parameter, which is supported by the computational implementation, allowing the generation of networks within a user-defined range of k. Intriguingly, the choice of the parameter allows for the effective investigation of admixture between populations at larger values of k (connecting more distantly related individuals), whereas smaller values of the parameter allow for the examination of fine-scale genetic sub-structure (connecting more closely related individuals) (Neuditschko *et al.* 2012). Therefore, the value of the parameters also depends on the questions that are asked about the data and may be applicable to a broad range of applications in population genetics. Furthermore, genetically distinct individuals are immediately recognisable, including anomalous samples away from their supposed family or population of origin (for instance, one wolf from Idaho linked to New Mexico). *NetView P* could therefore also be used for detecting unexpected genetic relatedness or incorrect sample assignments, knowledge of which may be useful for breeding programs or phylogenetic studies (Neuditschko *et al.* 2012).

It should be noted that the degree of divergence between groups is only approximated by the general location of clusters in the networks and the association of edge width and genetic distance. Additional investigations into the quantitative degree and statistical significance of differentiation or ecological exchangeability of populations may therefore be useful when considering problems such as the delineation of conservation management units (Crandall *et al.* 2000; Palsbøll *et al.* 2007). Finally, the original clustering algorithm SPC (Blatt *et al.* 1996) was replaced with the optional community detection algorithm Infomap (Rosvall & Bergstrom 2008) in order to provide an integrated and user-friendly implementation of the pipeline in Python. However, its application to genetic population data remains to be appropriately investigated. Objective methods for delineating communities in the

networks may be particularly useful when prior information about the populations (e.g. geographical location or pedigree records) is lacking. A discussion and implementation of the wide variety of algorithms available for this purpose (Girvan & Newman 2002; Pons & Latapy 2006; Rosvall & Bergstrom 2008; Ahn *et al.* 2010; Rodriguez & Laio 2014) was, however, beyond the scope of this study. Regardless, it is often of more immediate interest to relate available population information (e.g. geographical location, pre-defined populations or pedigree records) to the genetic structure that emerges from the data. Here, this structure is determined solely from genetic similarities and supplemented with meta-data that can be readily visualized in the network topologies generated by *NetView P*.

As the acquisition of high-quality population genomics data is becoming increasingly cost-effective, population structure analyses can now be carried out for hundreds of individuals, in both model and non-model species, with thousands to hundreds of thousands of high-quality SNPs. *NetView P* provides a network-based addition to model-based approaches of population analysis with the potential to reveal large- and fine-scale patterns of genetic structure in wild and captive populations. The implementation of the pipeline is computationally efficient through multiprocessing capabilities and can generate high-definition visualisations of complex genetic structure. In addition to other available methods of network analysis, such as *popgraph* for R (Dyer & Nason 2004) or EDENetworks (Kivelä *et al.* 2014) the integrated pipeline is now accessible in Python and can be used for the exploration and visualization of population structure derived from genome-wide SNPs.

Acknowledgements

We would like to express our sincere gratitude to Monal Lal and Shannon Kjeldsen for testing *NetView P*, to David Jones and Curtis Lind for support with the data from *P.maxima* and to the two anonymous reviewers who provided constructive comments on the manuscript.

References

Ahn Y-Y, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–4.

Arnaud-Haond S, Vonau V, Rouxel C *et al.* (2008) Genetic structure at different spatial scales in the pearl oyster (Pinctada margaritifera cumingii) in French Polynesian lagoons: beware of sampling strategy and genetic patchiness. *Marine Biology*, **155**, 147–157.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, **3**, e3376.

Barad O (2003) Advanced clustering algorithm for the analysis of gene expression data. *MSc Thesis, Weizmann Institute, Israel*.

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Batagelj V, Mrvar A (1998) Pajek- Program for Large Network Analysis. *Connections*, **21**, 47–57.

Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. *Physical Review Letters*, **76**, 3251–3254.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–40.

Crandall K, Bininda-Emonds O, Mace G, Wayne R (2000) Considering evolutionary processes in conservation biology. *Trends in Ecology & Evolution*, **15**, 290–295.

Cronin MA, Cánovas A, Islas-Trejo A *et al.* (2015) Single nucleotide polymorphism loci (SNP) variation of wolves (Canis lupus) in Southeast Alaska and comparison with wolves and coyotes in North America. *Journal of Heredity*, **106**, 26–36.

Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695.

Doncheva NT, Assenov Y, Domingues FS, Albrecht M (2012) Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, **7**, 670–85.

Dyer RJ, Nason JD (2004) Population Graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, **13**, 1713–27.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Frean M, Rainey PB, Traulsen A (2013) The effect of population structure on the rate of evolution. *Proceedings of the Royal Society B: Biological Sciences* , **280** .

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7821–6.

Hayes B, He J, Moen T, Bennewitz J (2006) Use of molecular markers to maximise diversity of founder populations for aquaculture breeding programs. *Aquaculture*, **255**, 573–578.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews. Genetics*, **10**, 639–650.

Ivy JA, Lacy RC (2012) A Comparison of Strategies for Selecting Breeding Pairs to Maximize Genetic Diversity Retention in Managed Populations. *Journal of Heredity* , **103** , 186–196.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Jones DB, Jerry DR, Forêt S, Konovalov DA, Zenger KR (2013a) Genome-wide SNP validation and mantle tissue transcriptome analysis in the silver-lipped pearl oyster, Pinctada maxima. *Marine Biotechnology (New York, N.Y.)*, **15**, 647–58.

Jones DB, Jerry DR, Khatkar MS, Raadsma HW, Zenger KR (2013b) A high-density SNP genetic linkage map for the silver-lipped pearl oyster, Pinctada maxima: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics*, **14**, 810.

Kivelä M, Arnaud-Haond S, Saramäki J (2014) EDENetworks: A user-friendly software to build and analyse networks in biogeography, ecology and population genetics. *Molecular Ecology Resources*.

Lind CE, Evans BS, Knauer J, Taylor JJU, Jerry DR (2009) Decreased genetic diversity and a reduced effective population size in cultured silver-lipped pearl oysters (Pinctada maxima). *Aquaculture*, **286**, 12–19.

Lind CE, Evans BS, Taylor JJU, Jerry DR (2007) Population genetics of a marine bivalve, Pinctada maxima, throughout the Indo Australian Archipelago shows differentiation and decreased diversity at range limits. *Molecular Ecology*, **16**, 5193–5203.

Lind CE, Ponzoni RW, Nguyen NH, Khaw HL (2012) Selective breeding in fish and conservation of genetic resources for aquaculture. *Reproduction in Domestic Animals*, **47**, 255–63.

Maier M, Hein M, von Luxburg U (2007) Cluster Identification in Nearest-Neighbor Graphs. In: *Algorithmic Learning Theory SE - 18* Lecture Notes in Computer Science. (eds Hutter M, Servedio R, Takimoto E), pp. 196–210. Springer Berlin Heidelberg.

Milano I, Babbucci M, Cariani A *et al.* (2014) Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius). *Molecular Ecology*, **23**, 118–35.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–8.

Moalic Y, Arnaud-Haond S, Perrin C, Pearson GA, Serrao EA (2011) Travelling in time with networks: Revealing present day hybridization versus ancestral polymorphism between two species of brown algae, Fucus vesiculosus and F. spiralis. *BMC Evolutionary Biology*, **11**, 33.

Moalic Y, Desbruyères D, Duarte C *et al.* (2012) Biogeography revisited with network theory: retracing the history of hydrothermal vent communities. *Systematic Biology*, **61**, 127–137.

Neuditschko M, Khatkar MS, Raadsma HW (2012) NetView: a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. (NJ Timpson, Ed,). *PloS One*, **7**, e48375.

Noutsos C, Borevitz JO, Hodges SA (2014) Gene flow between nascent species: geographic, genotypic and phenotypic differentiation within and between Aquilegia formosa and A. pubescens. *Molecular Ecology*.

Palsbøll PJ, Bérubé M, Allendorf FW (2007) Identification of management units using population genetic data. *Trends in Ecology & Evolution*, **22**, 11–6.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Pons P, Latapy M (2006) Computing communities in large networks using random walks. *Journal of Graph Algorithms Applications*, **10**, 191–218.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–9.

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, **11**, 459–63.

Prim RC (1957) Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, **36**, 1389–1401.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* , **155** , 945–959.

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–28.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Rodriguez A, Laio A (2014) Machine learning. Clustering by fast search and find of density peaks. *Science (New York, N.Y.)*, **344**, 1492–6.

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1118–23.

Rozenfeld AF, Arnaud-Haond S, Hernández-García E *et al.* (2008) Network analysis identifies weak and strong links in a metapopulation system. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 18824–9.

Russello M, Amato G (2004) Ex situ population management in the absence of pedigree information. *Molecular Ecology*, **13**, 2829–2840.

Sargolzaei M, Schenkel FS (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics* , **25** , 680–681.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Tsafrir D, Tsafrir I, Ein-Dor L *et al.* (2005) Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics* , **21** , 2301–2308.

vonHoldt BM, Pollinger JP, Earl DA *et al.* (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, **21**, 1294–305.

Data Accessibility

NetView P

The computational implementation is freely available at: https://sourceforge.net/projects/netview-genomics/ . The distribution includes a detailed manual, example data and binaries for Infomap (Linux) and PLINK v1.07 (Linux, Windows).

Data Sets

Pinctada maxima: http://dx.doi.org/10.5061/dryad.p3b3f

Merluccius merluccius: http://dx.doi.org/10.5061/dryad.7bn22

Canis lupus: http://dx.doi.org/10.5061/dryad.284tf

Author Contributions

EJS assembled the data, performed the analyses, wrote the computational implementation and drafted the manual and manuscript, MN supported the analyses and drafted the manuscript, MSK carried out the simulations and drafted the manuscript, HWR conceived and coordinated the study and edited the manuscript, KRZ conceived, designed and coordinated the study, provided data and drafted the manuscript.

Figure Legends

Figure 1. Workflow for *NetView P*. Dashed boxes denote optional processes and input files.

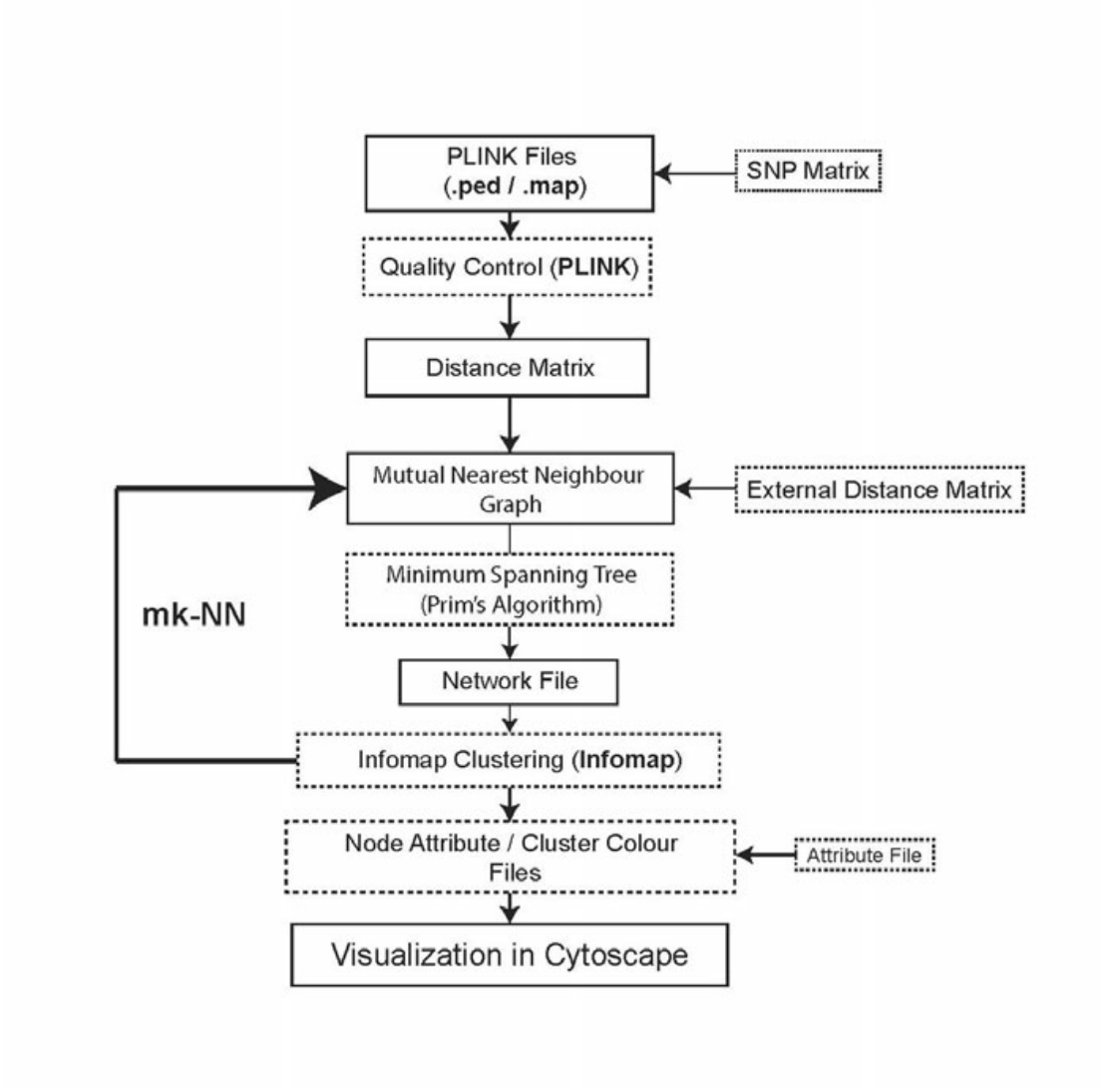
Figure 2. Organic network of *M. merluccius* (n = 849) at *k* = 10 (with MST), based on 380 SNPs. Colour shades represent sampling locations within the Atlantic (green) and the Mediterranean (blue), edge width is proportional to the genetic distance between individuals. Several genetically distinct individuals are connected by single edges at the periphery of the network and orange nodes highlight a close relationship of Algerian samples with the Atlantic.

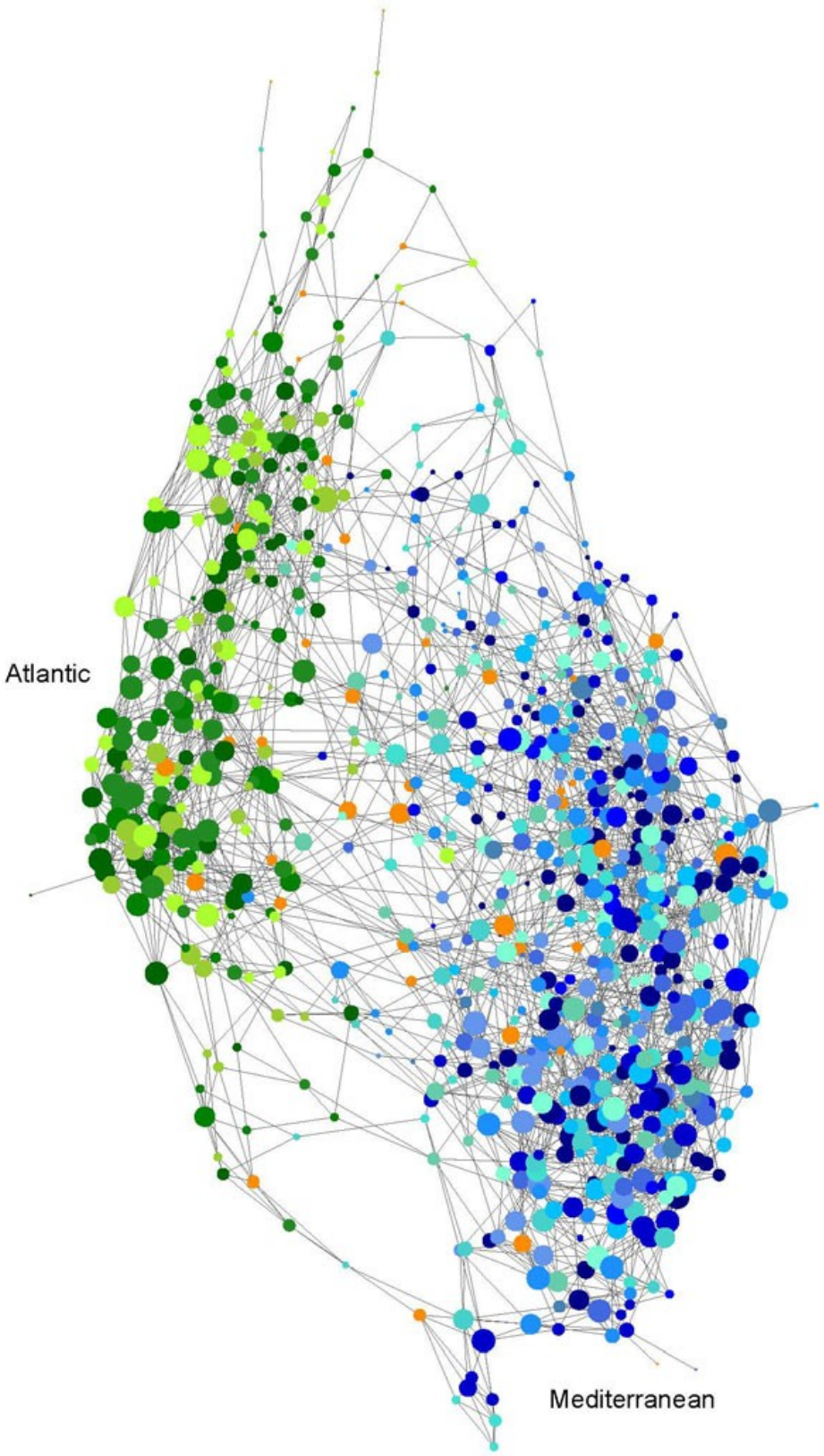
Figure 3. Organic network of *C. lupus* from North America (n = 239) at *k* = 10 (with MST), based on 86,103 SNPs. Edge width represents the genetic distance between individuals. Colours and labels denote sampling sites, including three Game Management Units (GMUs) in Southeast Alaska. The network visualisation clearly shows connectivity of individuals between populations and fine-scale genetic structure in populations from Alaska.

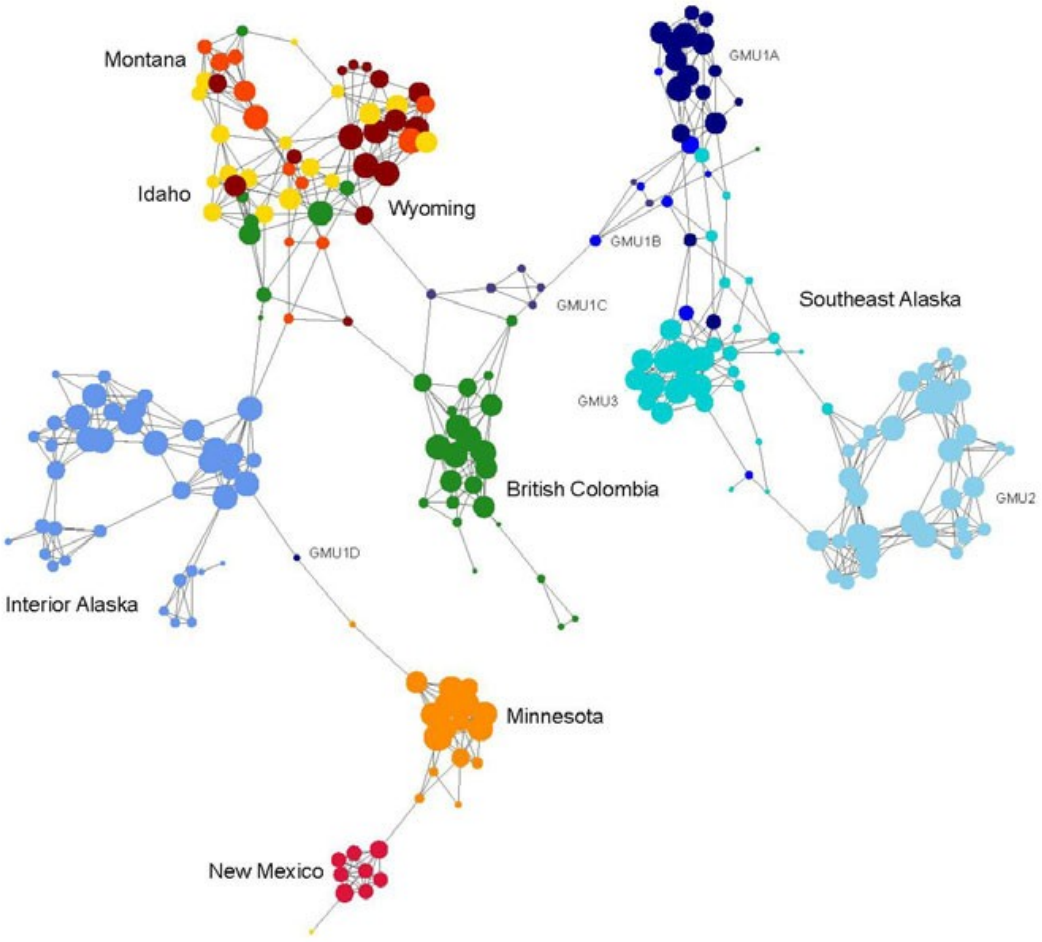
Figure 4. Organic network of simulated data (n = 750) at *k* = 10 (without MST) based on 1279 SNPs) (**A**) and family pedigree from data simulation with QMSim (**B**). Simulations were derived from a historic founder population (1000 generations) with random selection of 20 males and 20 females producing 50 offspring each over 10 discrete generations. Genotypes were generated for the last three generations (F8, F9, F10) and three families were randomly selected in F10, including their parental families in F8 and F9. Colours represent generation (purple: F8, green: F9, orange: F10). Individuals are accurately assigned to their respective families and unconnected networks have been ordered manually to represent the pedigree from the simulation.

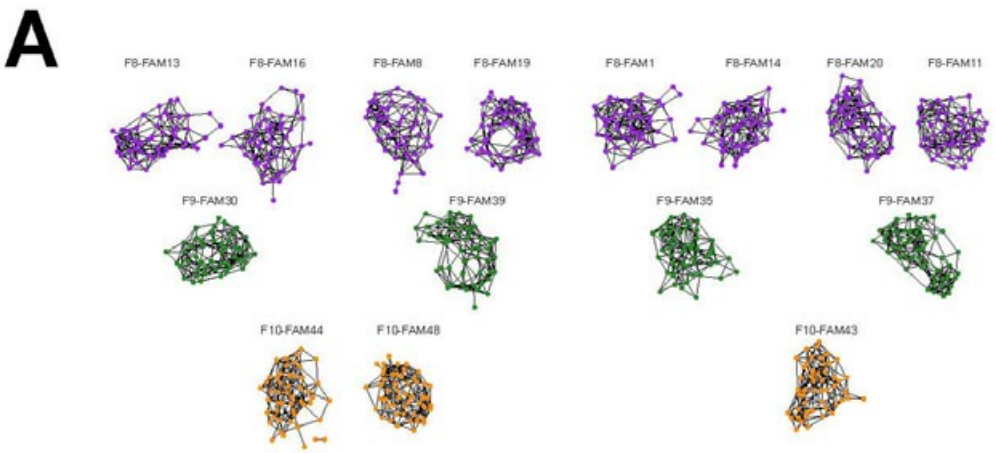
Figure 5. Circular network of simulated data (n = 750) at *k* = 10 (with MST) based on 1279 SNPs. The general placement of the individual families reflects the pedigree of the data, with edges from the MST (red) connecting most parents (darkred) to their offspring families in F9 and F10. Node colours reflect generation (purple: F8, green: F9, orange: F10) and families and parents are labelled according to the simulated pedigree.

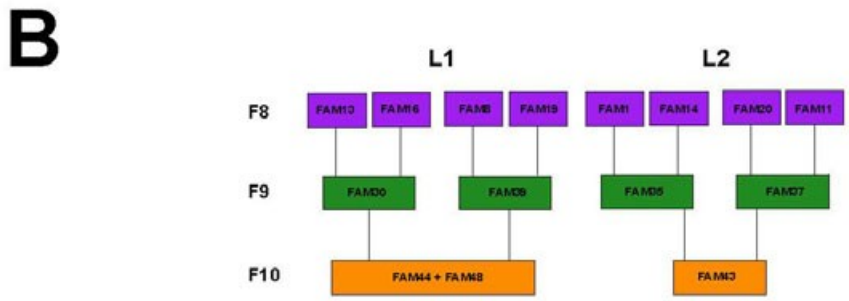
Figure 6. Circular network of captive populations of *P. maxima* in F_0 to F_2 (n = 415) at *k* = 10 (with MST), based on 1107 SNPs. Node colours and labels depicts family assignment of individuals based on previously assembled pedigree records. Families and relationship between families and individuals are clearly visible in the network topology. Asteriks (*) denote half-sibling relationships between F1-BB-1 and F1-BB-2, as well as F1-BB-3 and F1-BB-4. Triangular nodes denote parents of F_0 and rectangular nodes denote parents of F_1.Red edges are derived from the MST.

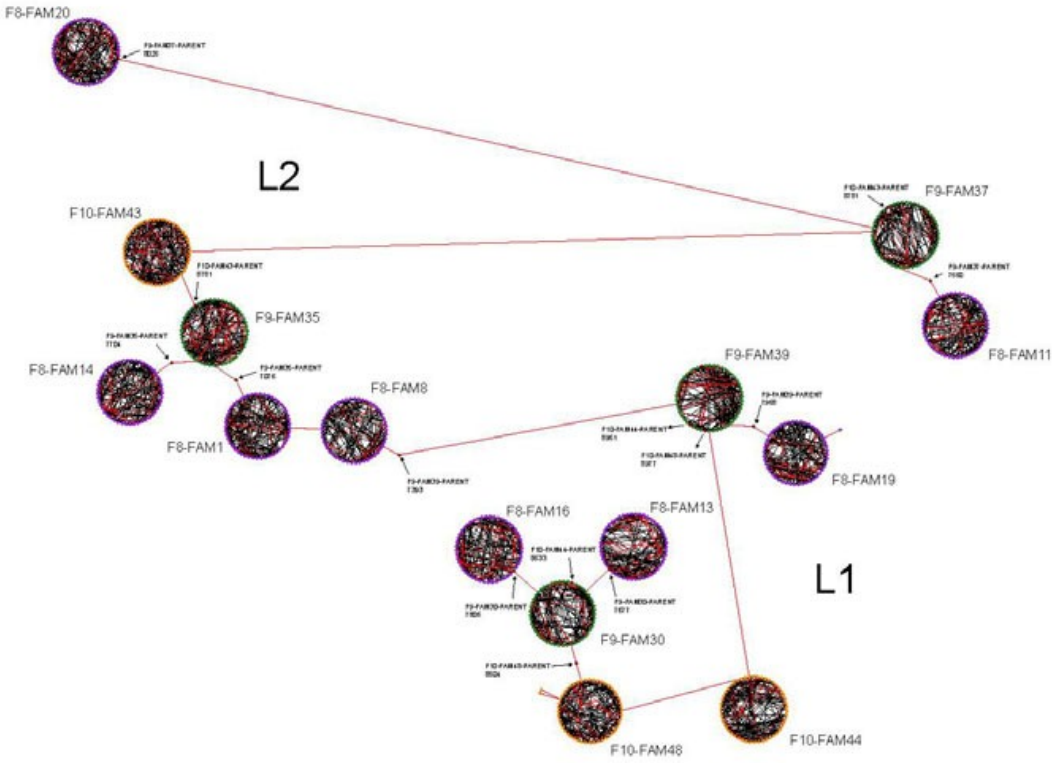

PLINK Files
(.ped / .map)
SNP Matrix
Quality Control (PLINK)
Distance Matrix
Mutual Nearest Neighbour Graph
External Distance Matrix
Minimum Spanning Tree (Prim's Algorithm)
mk-NN
Network File
Infomap Clustering (Infomap)
Node Attribute / Cluster Colour Files
Attribute File
Visualization in Cytoscape

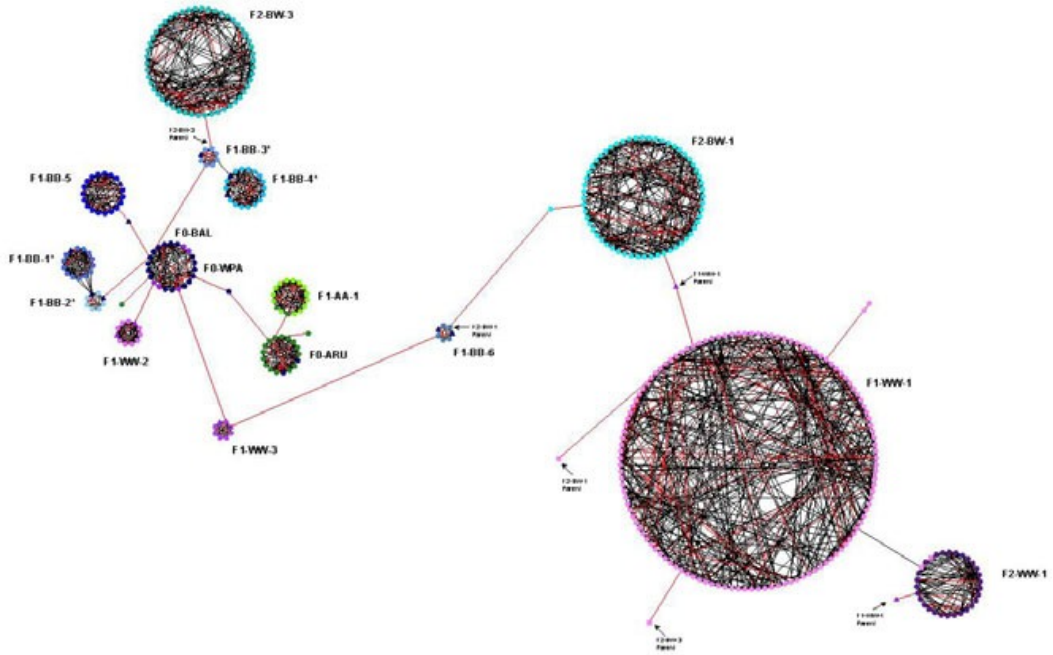

Atlantic
Mediterranean


Montana
Idaho
Wyoming
GMU1A
GMU1B
GMU1C
Southeast Alaska
GMU3
GMU2
British Colombia
GMU1D
Interior Alaska
Minnesota
New Mexico


A
F8-FAM13
F8-FAM16
F8-FAM8
F8-FAM19
F8-FAM1
F8-FAM14
F8-FAM20
F8-FAM11
F9-FAM30
F9-FAM39
F9-FAM35
F9-FAM37
F10-FAM44
F10-FAM48
F10-FAM43


B
L1
L2
F8
F9
F10
FAM13
FAM16
FAM8
FAM19
FAM1
FAM14
FAM20
FAM11
FAM30
FAM39
FAM35
FAM37
FAM44 + FAM48
FAM43


F8-FAM20
L2
F10-FAM43
F9-FAM37
F9-FAM35
F8-FAM11
F8-FAM14
F8-FAM8
F9-FAM39
F8-FAM1
F8-FAM19
F8-FAM16
F8-FAM13
L1
F9-FAM30
F10-FAM48
F10-FAM44


F2-BW-3
F1-BB-3*
F1-BB-4*
F1-BB-5
F0-BAL
F1-BB-1*
F0-WPA
F1-BB-2*
F1-AA-1
F1-WW-2
F0-ARU
F1-BB-6
F1-WW-3
F2-BW-1
F1-WW-1
F2-WW-1